Code Generation: Results Achieved with Zero-Shot prompting

Table 1: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

LLM	Code generation			
	Java		Python	
	#	%	#	%
DeepSeek Coder 1.3B	635	45.20%	137	10.69%
DeepSeek Coder 6.7B	19	1.35%	13	1.01%
DeepSeek Coder 33B	4	0.28%	134	10.46%
CodeLlama 7B	7	0.07%	138	10.77%
CodeLlama 13B	3	0.21%	23	1.80%
CodeLlama 34B	2	0.14%	1	0.08%
GPT-3.5-turbo	0	0.00%	6	0.47%
GPT-4-turbo	25	1.78%	0	0.00%

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.72	0.01	-0.25	-0.04	0.53	0.55	0.48	0.59	-0.86	*** (M)	*** (S)	*** (L)
DSC 6.7B	0.78	0.76	0.64	0.76	0.62	0.67	0.65	0.63	-0.01	* (N)	* (N)	*** (L)
DSC 33B	0.26	0.29	0.23	0.26	0.22	0.20	0.32	0.29	-0.69	(N)	(N)	*** (L)
CL 7B	-0.21	-0.24	-0.34	-0.22	-0.36	-0.29	-0.33	-0.35	-1.00	(N)	(N)	*** (L)
CL 13B	-0.03	-0.15	-0.23	-0.11	-0.20	-0.19	0.21	0.02	-0.81	** (N)	** (S)	*** (L)
CL 34B	0.24	0.26	0.16	0.25	0.17	0.19	0.39	0.37	-0.47	(N)	(N)	*** (L)
GPT-3.5	0.34	0.56	0.43	0.54	0.35	0.34	0.48	0.48	-0.29	(N)	(N)	*** (L)
GPT-4	0.34	0.42	0.39	0.37	0.34	0.30	0.48	0.52	-0.37	** (N)	** (N)	*** (L)
Average (all)	**0.23**	**0.28**	**0.20**	**0.26**	**0.17**	**0.17**	**0.38**	**0.34**	**-0.53**	-	-	-
Average (large)	**0.30**	**0.24**	**0.13**	**0.23**	**0.21**	**0.22**	**0.34**	**0.32**	**-0.56**	**-**	**-**	**-**

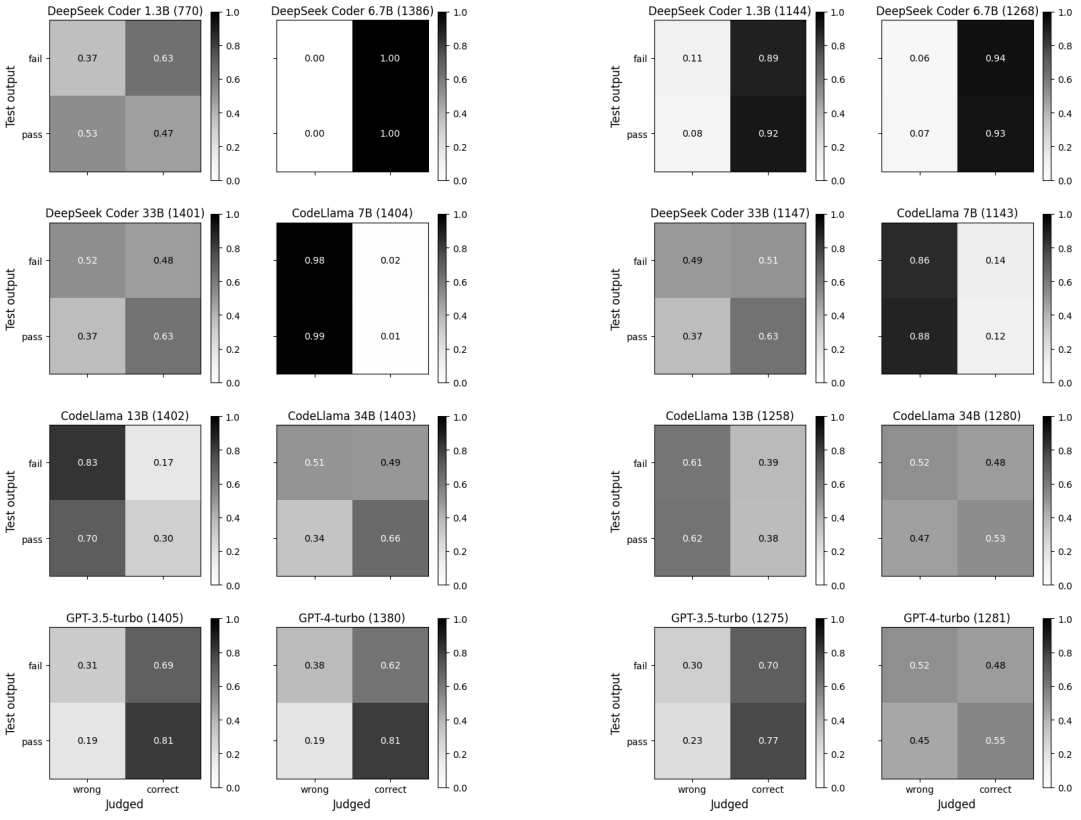
Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 2: **(Java)** Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.75	0.74	0.77	0.76	0.60	0.64	0.60	0.69	-0.10	(N)	(N)	*** (L)
DSC 6.7B	0.79	0.80	0.77	0.79	0.61	0.73	0.69	0.74	-0.12	(N)	(N)	*** (L)
DSC 33B	0.30	0.37	0.36	0.39	0.30	0.33	0.30	0.46	-0.49	(N)	(N)	*** (L)
CL 7B	-0.05	-0.06	0.00	0.00	-0.12	-0.09	-0.17	-0.02	-0.89	(N)	(N)	*** (L)
CL 13B	0.24	0.27	0.22	0.24	0.10	0.13	0.15	0.12	-0.66	(N)	(N)	*** (L)
CL 34B	0.36	0.33	0.32	0.39	0.22	0.19	0.17	0.27	-0.45	(N)	(N)	*** (L)
GPT-3.5	0.48	0.57	0.52	0.50	0.42	0.49	0.58	0.65	-0.32	(N)	(N)	*** (L)
GPT-4	0.21	0.34	0.34	0.18	0.23	0.23	0.42	0.65	-0.66	*** (M)	*** (M)	*** (L)
Average (all)	**0.32**	**0.38**	**0.35**	**0.34**	**0.25**	**0.27**	**0.32**	**0.43**	**-0.52**	-	-	-
Average (large)	**0.38**	**0.42**	**0.41**	**0.41**	**0.30**	**0.33**	**0.34**	**0.44**	**-0.46**	**-**	**-**	**-**

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 3: **(Python)** Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.



(a) Java (b) Python

Figure 1: Code Generation: Confusion matrices for LLM's judgment.

Code Generation: Results Achieved with Zero-Shot prompting when NOT asking for a Rationale

Table 4: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

LLM	Code generation			
	Java		Python	
	#	%	#	%
DeepSeek Coder 1.3B	379	26.98%	31	2.42%
DeepSeek Coder 6.7B	0	0.00%	3	0.23%
DeepSeek Coder 33B	3	0.21%	3	0.23%
CodeLlama 7B	1	0.07%	61	4.76%
CodeLlama 13B	3	0.21%	44	3.43%
CodeLlama 34B	3	0.21%	9	0.70%
GPT-3.5-turbo	0	0.00%	0	0.00%
GPT-4-turbo	0	0.00%	0	0.00%

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.64	0.58	0.47	0.57	0.47	0.54	0.49	0.57	-0.11	(N)	(N)	*** (L)
DSC 6.7B	0.78	0.76	0.65	0.76	0.63	0.68	0.65	0.64	-0.01	* (N)	* (N)	*** (L)
DSC 33B	0.35	0.42	0.30	0.39	0.32	0.35	0.37	0.40	-0.60	(N)	(N)	*** (L)
CL 7B	-0.12	-0.18	-0.24	-0.13	-0.28	-0.20	-0.14	-0.24	-0.93	(N)	(N)	*** (L)
CL 13B	0.62	0.56	0.43	0.61	0.46	0.49	0.61	0.57	-0.09	* (N)	* (N)	*** (L)
CL 34B	0.23	0.22	0.12	0.21	0.17	0.13	0.32	0.30	-0.52	(N)	* (N)	*** (L)
GPT-3.5	0.53	0.62	0.48	0.59	0.51	0.55	0.57	0.55	-0.23	(N)	(N)	*** (L)
GPT-4	0.34	0.44	0.34	0.38	0.35	0.33	0.49	0.52	-0.42	** (N)	** (N)	*** (L)
Average (all)	**0.42**	**0.45**	**0.34**	**0.44**	**0.36**	**0.37**	**0.47**	**0.47**	**-0.37**	-	-	-
Average (large)	**0.42**	**0.43**	**0.32**	**0.42**	**0.33**	**0.36**	**0.42**	**0.41**	**-0.36**	**-**	**-**	**-**

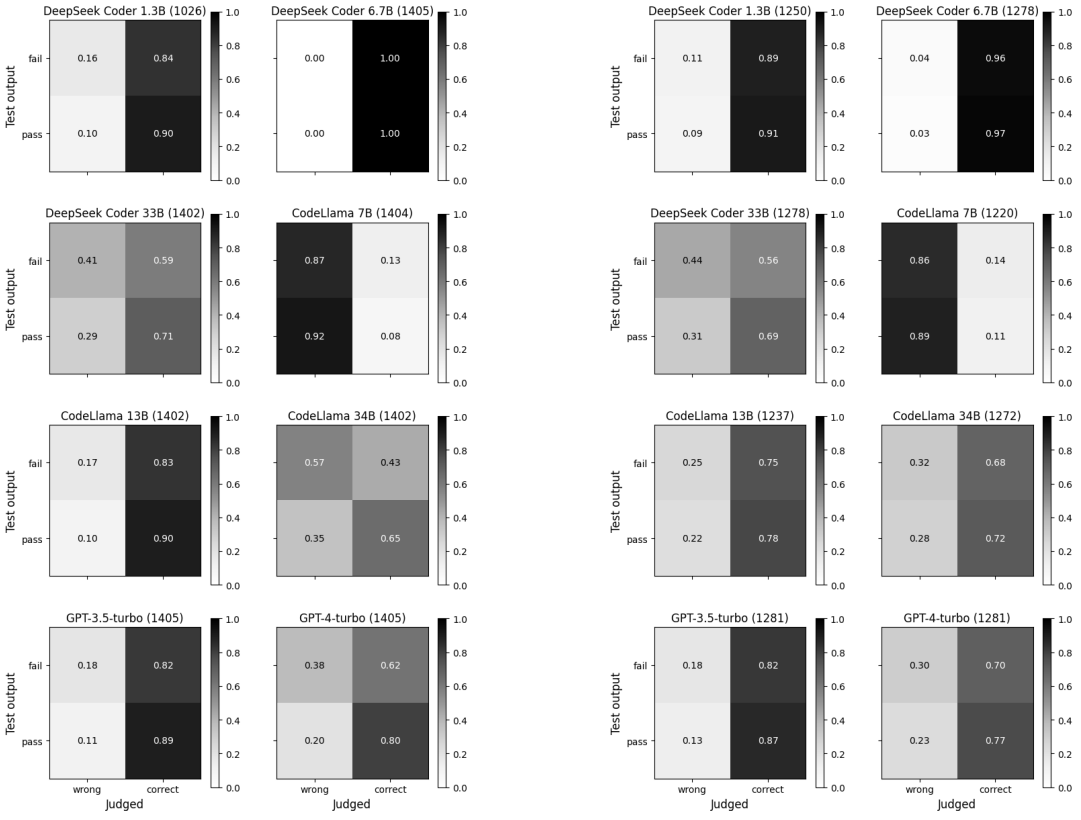
Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 5: **(Java)** Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.78	0.72	0.69	0.76	0.60	0.71	0.65	0.63	-0.11	* (N)	* (N)	*** (L)
DSC 6.7B	0.81	0.81	0.79	0.78	0.68	0.73	0.72	0.74	-0.03	(N)	(N)	*** (L)
DSC 33B	0.31	0.46	0.42	0.38	0.38	0.34	0.40	0.53	-0.45	(N)	(N)	*** (L)
CL 7B	0.01	-0.01	-0.08	0.07	-0.17	-0.14	-0.13	-0.04	-0.91	** (N)	* (N)	*** (L)
CL 13B	0.59	0.59	0.62	0.60	0.52	0.55	0.43	0.52	-0.23	(N)	(N)	*** (L)
CL 34B	0.47	0.48	0.52	0.52	0.46	0.47	0.45	0.52	-0.30	(N)	(N)	*** (L)
GPT-3.5	0.64	0.71	0.64	0.63	0.55	0.63	0.63	0.69	-0.22	(N)	(N)	*** (L)
GPT-4	0.38	0.59	0.58	0.48	0.43	0.48	0.58	0.71	-0.36	*** (S)	*** (S)	*** (L)
Average (all)	**0.48**	**0.56**	**0.56**	**0.52**	**0.47**	**0.49**	**0.50**	**0.59**	**-0.31**	-	-	-
Average (large)	**0.50**	**0.54**	**0.52**	**0.52**	**0.43**	**0.47**	**0.47**	**0.54**	**-0.33**	**-**	**-**	**-**

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 6: **(Python)** Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.



(a) Java (b) Python

Figure 2: Code Generation: Confusion matrices for LLM's judgment.

Code Generation: Results Achieved with "Slow-Thinking" prompting

Table 7: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment.

LLM	Code generation			
	Java		Python	
	#	%	#	%
DeepSeek Coder 1.3B	143	10.18%	218	17.02%
DeepSeek Coder 6.7B	194	13.81%	131	10.23%
DeepSeek Coder 33B	34	2.42%	35	2.73%
CodeLlama 7B	98	6.98%	168	13.11%
CodeLlama 13B	352	25.05%	285	22.25%
CodeLlama 34B	68	4.84%	195	15.22%
GPT-3.5-turbo	4	0.28%	0	0.00%
GPT-4-turbo	0	0.00%	0	0.00%

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.65	0.62	0.53	0.66	0.59	0.54	0.52	0.56	-0.16	(N)	(N)	*** (L)
DSC 6.7B	0.67	0.55	0.47	0.63	0.46	0.51	0.52	0.48	-0.23	(N)	(N)	*** (L)
DSC 33B	0.56	0.67	0.54	0.62	0.51	0.54	0.58	0.57	-0.10	(N)	(N)	*** (L)
CL 7B	0.27	0.16	0.08	0.25	0.11	0.22	0.20	0.04	-0.62	(N)	(N)	*** (L)
CL 13B	0.65	0.52	0.37	0.65	0.44	0.55	0.56	0.49	-0.19	(N)	(N)	*** (L)
CL 34B	0.42	0.36	0.30	0.41	0.22	0.35	0.40	0.35	-0.35	(N)	(N)	*** (L)
GPT-3.5	0.54	0.66	0.51	0.65	0.53	0.53	0.61	0.62	-0.11	(N)	(N)	*** (L)
GPT-4	0.37	0.45	0.38	0.43	0.37	0.35	0.48	0.53	-0.25	* (N)	* (N)	*** (L)
Average (all)	**0.51**	**0.53**	**0.42**	**0.55**	**0.41**	**0.47**	**0.52**	**0.51**	**-0.20**	-	-	-
Average (large)	**0.52**	**0.50**	**0.40**	**0.54**	**0.40**	**0.45**	**0.48**	**0.46**	**-0.25**	**-**	**-**	**-**

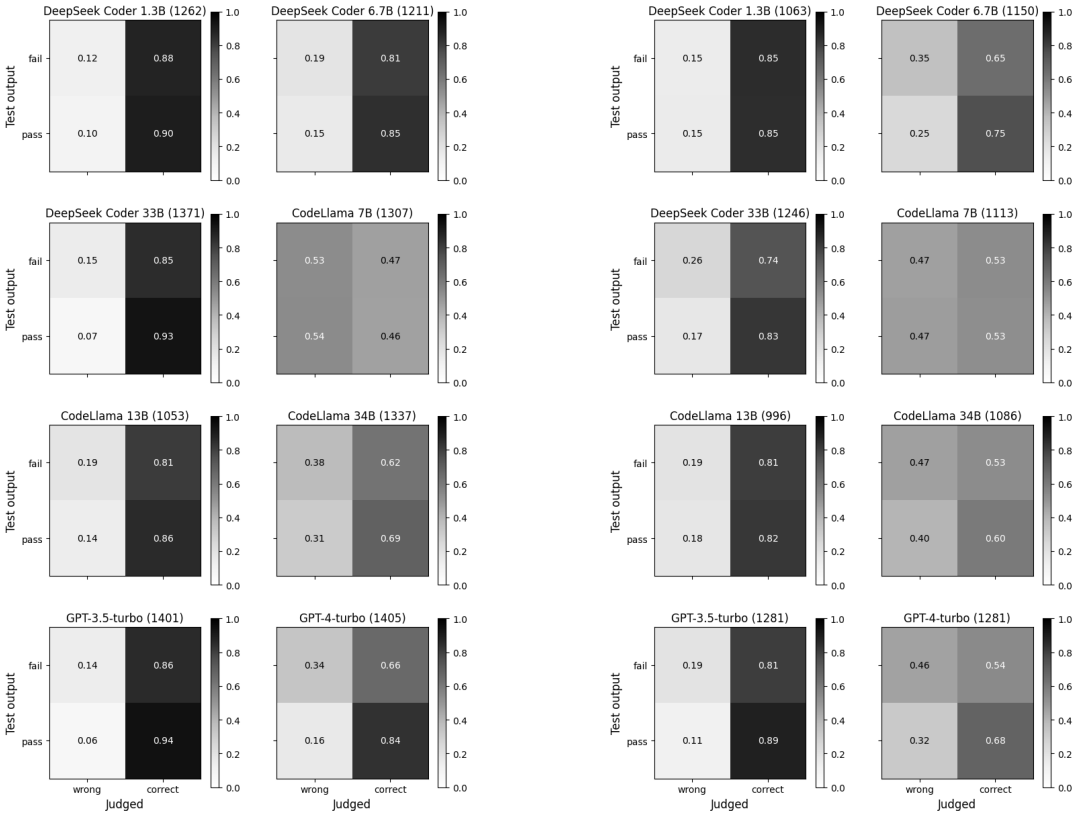
Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 8: **(Java)** Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.71	0.71	0.69	0.71	0.54	0.66	0.62	0.64	-0.22	(N)	(N)	*** (L)
DSC 6.7B	0.44	0.55	0.50	0.49	0.36	0.47	0.39	0.50	-0.27	(N)	(N)	*** (L)
DSC 33B	0.46	0.67	0.63	0.54	0.45	0.57	0.57	0.58	-0.19	(N)	(N)	*** (L)
CL 7B	0.32	0.43	0.37	0.31	0.19	0.32	0.25	0.34	-0.42	(N)	(N)	*** (L)
CL 13B	0.68	0.65	0.70	0.60	0.46	0.59	0.58	0.46	-0.16	* (N)	* (N)	*** (L)
CL 34B	0.37	0.48	0.35	0.33	0.29	0.32	0.33	0.29	-0.41	(N)	(N)	*** (L)
GPT-3.5	0.61	0.64	0.65	0.57	0.53	0.57	0.63	0.73	-0.11	(N)	(N)	*** (L)
GPT-4	0.20	0.39	0.36	0.37	0.31	0.34	0.40	0.64	-0.38	*** (S)	*** (S)	*** (L)
Average (all)	**0.46**	**0.57**	**0.54**	**0.48**	**0.41**	**0.48**	**0.50**	**0.54**	**-0.25**	-	-	-
Average (large)	**0.47**	**0.56**	**0.53**	**0.49**	**0.39**	**0.48**	**0.47**	**0.52**	**-0.27**	**-**	**-**	**-**

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 9: **(Python)** Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.



(a) Java (b) Python

Figure 3: Code Generation: Confusion matrices for LLM's judgment.